Comparing Dynamics: Deep Neural Networks versus Glassy Systems

Marco Baity-Jesi 1  Levent Sagun 2,3  Mario Geiger 3  Stefano Spigler 3,2  Gérard Ben Arous 4  
Chiara Cammarota 5  Yann LeCun 4,6,7  Matthieu Wyart 3  Giulio Biroli 2,8

Abstract

We analyze numerically the training dynamics of deep neural networks (DNN) by using methods developed in statistical physics of glassy systems. The two main issues we address are (1) the complexity of the loss landscape and of the dynamics within it, and (2) to what extent DNNs share similarities with glassy systems. Our findings, obtained for different architectures and datasets, suggest that during the training process the dynamics slows down because of an increasingly large number of flat directions. At large times, when the loss is approaching zero, the system diffuses at the bottom of the landscape. Despite some similarities with the dynamics of mean-field glassy systems, in particular, the absence of barrier crossing, we find distinctive dynamical behaviors in the two cases, showing that the statistical properties of the corresponding loss and energy landscapes are different. In contrast, when the network is under-parametrized we observe a typical glassy behavior, thus suggesting the existence of different phases depending on whether the network is under-parametrized or over-parametrized.

1. Introduction

The training process of a deep neural network (DNN) shares very strong similarities with the physical dynamics of disordered systems: the loss function plays the role of the energy, the weights are the degrees of freedom, and the dataset corresponds to the parameters defining the energy function. The randomness in the data is akin to what is called “quenched disorder” in the physics literature. Training is routinely performed by the Stochastic Gradient Descent (SGD), which consists in starting from random initial conditions and then letting the weights evolve dynamically towards configurations corresponding to low loss values. This process is, in fact, similar to what is called “a quench” in physics. The quenching protocol corresponds to a sudden decrease of the thermal noise, usually done by lowering the temperature of the thermal bath, for a system which is initially prepared in equilibrium at very high temperature. The study of the dynamics induced by quenches has been one of the most important topics of out-of-equilibrium physics of the last decades (Biroli, 2016). The main model considered in the literature is based on stochastic Langevin equations, reminiscent of SGD and corresponding to an evolution governed by gradient descent plus random noise. Since the initial temperature is very high, the initial conditions for the dynamics are random, featureless and uncorrelated with the quenched disorder if present, again in strong analogy with DNNs. Disordered systems are known to display glassy dynamics after a quench, which means that the system gets stuck for long times in local minima (Biroli, 2016; Bouchaud et al., 1998; Berthier & Biroli, 2011; Cugliandolo, 2003). Given the similarity between the training of DNNs and quenching of disordered systems, it may seem surprising that meaningful local minima with perfect accuracy on the training set are found (Zhang et al., 2016).

In the current literature, several explanations are proposed to explain this paradox. Two quite different points of view emerge from it. One is that even though the loss function displays a very large number of local minima with different loss values, the dynamics during the training process allows the system to decrease the loss without barrier crossing and to converge towards quite low local minima that allow good generalization. In other words, the loss landscape is very rough, however, this doesn’t damage the performance of...
the system. In this direction, (Choromanska et al., 2015) proposed an analogy with mean-field glassy systems. In such systems, it was shown by theoretical physics methods (Cugliandolo & Kurchan, 1993), backed up by rigorous results (Ben Arous et al., 2006), that dynamics corresponding to gradient descent or stochastic versions of it tend without barrier crossing to the widest and the highest minima, despite the existence of deeper local and global minima. A complementary point of view, proposed in (Baldassi et al., 2016), is that there exist rare and wide minima which have large basins of attraction and are reached without any substantial barrier crossing by the training dynamics.

Another quite different point of view is that deep neural networks work in a regime in which there are actually no spurious local minima that can trap the system during the training process. Several rigorous and numerical works, including but not limited to (Freeman & Bruna, 2016; Hoffer et al., 2017; Soudry & Carmon, 2016), suggest that the loss function, despite being non-convex, is characterized by a connected level set as long as one considers loss values above the global minimum. From this perspective, the dynamical evolution induced by the stochastic gradient descent corresponds to falling down in the loss landscape without barrier crossing. In this case, it is the absence of bad local minima, and consequently, the absence of roughness and glassy dynamics, that solves the previous paradox.

Beyond the above two seemingly contradictory pictures on the structure of the loss landscape, there is also a rich literature discussing the path the dynamical process takes during the training process. For instance, Dauphin et al. (2014) claims that it is the existence of numerous saddle points that lie on the dynamical paths that present itself as a form of an obstacle to find deeper local minimum. Several other works, including Lee et al. (2016), claim that gradient-based training avoids such obstacles even if they do exist. And finally, Lipton (2016) demonstrates how the weights travel large distances through the flat basins by looking at the principle components of the evolution of the weights.

Establishing conclusively these scenarios in realistic cases is a challenge. Exact calculations of the statistical properties of critical points are hampered by the increased computational complexity of over-parametrized models and the possible degeneracy of critical points. Some guidance is provided by empirical results. In fact, simulations in Sagun et al. (2014) demonstrate that different dynamical processes on the loss landscape can indeed perform similarly regardless of the effect of the noise of SGD, thus suggesting that barrier crossing indeed does not take place. The works Keskar et al. (2016) and Jastrzbski et al. (2017) claim that by tuning the hyper-parameters of the system one can locate local minima with different qualities, thus providing indications of the roughness of the loss landscape. The results of Chaudhari et al. (2016) demonstrate that wider and possibly rarer basins can be found by averaging out the values of several parallel optimizers.

At the moment, it is still not clear what approach provides a good answer. It could be actually that the correct one contains ingredients from all the perspectives cited above. In this work, we address this problem by taking advantage of knowledge gained in the field of glassy out-of-equilibrium systems in the last decades (Bray, 2002; Biroli, 2016; Bouchaud et al., 1998). Our approach is twofold: (1) probing the training dynamics through the measurement of one and two-point correlation functions, as done in physics, we infer properties of the loss landscape in which the system is evolving, (2) comparing the results obtained for mean-field glasses to measurements performed for realistic DNNs we test the analogy between these systems.

Our Contribution: The analysis is performed for several different architectures, see Sec. 3, varying from specific toy models to ResNets (He et al., 2016) which are evaluated on popular datasets such as MNIST and CIFAR. We decided to focus both on a simple architecture and on more competitive ones. The former is close to a model where, for a large-enough hidden layer, there is a proof of the non-existence of bad local minima (Freeman & Bruna, 2016), and the latter are a relatively more realistic one with relevant performances on the given task. The dynamical behavior we found is similar for all cases: After an initial exploration of high-loss configurations, the system starts its descent in the “loss landscape”, and displays a particular kind of glassy dynamics, called aging, see Sec. 2. Our results suggest that the slowness of the dynamics in this stage is not related to the crossing of large barriers but instead to the emergence of an increasingly large number of flat directions (Sagun et al., 2017). At long times, a stationary regime where aging is interrupted and the system becomes almost stationary sets in. We present evidences that this dynamical regime corresponds to diffusion, not necessarily isotropic (as suggested by (Jastrzbski et al., 2017)), at or close to the bottom of the loss landscape. We compare these behaviors to the ones of the $p$-spin spherical model, which is one of the most studied mean-field glass models. We find that although the first regimes share similarities with the dynamics of mean-field glasses after a quench, the final regime does not. This suggests a qualitative different geometrical characterization of the bottom of the loss landscape and, accordingly, of the dynamics within it.

2. Basic facts on glassy dynamics

Two main observables have been identified as central to characterize the slow dynamics of physical systems. The first one is the energy as a function of time. When a system is quenched from high to low temperature the energy...
The mean square displacement of the p-spin model as a function of time in logarithmic scale after a sudden quench from a temperature $T_i = \infty$ to a temperature $T_f = 0.5$, for $p = 3$. In Figure 1(a), we also show an exponential decay, for comparison. In Figure 1(b), the mean-square-displacement is displayed for several $t_w$, increasing from left to right.

The existence of conserved quantities can produce a power-law dependence even in high-temperature phases. The two times are chosen in order to explicitly probe the out-of-equilibrium nature of the dynamics: $t_w$ is the time lapse after the quench, $t$ is the difference between the two times at which system configurations are compared. When the system is out-of-equilibrium, in particular after the quench, $\Delta(t_w, t_w + t)$ explicitly depends on both $t_w$ and $t$, whereas when equilibrium is reached the system becomes stationary and $\Delta(t_w, t_w + t)$ only depends on $t$. When quenched at low temperature many disordered systems show the phenomenon of aging, which means that the time-scale controlling the $t$-dependence is a function of $t_w$. In other words, the time it takes for the system to decorrelate depends on the age of the system.

In Figure 1(b), we plot $\Delta(t_w, t_w + t)$ for the 3-spin model as a function of $t$ and for different values of $t_w$. Focusing on the $t$-dependence, one can recognize the first time regime, which appears almost independent of $t_w$, in which the system appears stationary. This regime eventually ends at a time that increases with $t_w$. Then, the second regime which physically corresponds to aging emerges. Here, the longer is $t_w$ the longer it takes for the system to diffuse, i.e. for the large-time limit of $\Delta(t_w, t_w + t)$ is equal to two, as it should for diffusion on a sphere, where displacements are bounded.
mean-square displacement to escape from the plateau value. The height of the plateau is called Edwards-Anderson parameter in the physical literature and quantifies how much the system is frozen into a local minimum (Castellani & Cavagna, 2005).

Slow dynamics and aging are distinctive features of any glassy system. Particularly, in the p-spin spherical model, and in other models of glasses, the slow dynamics observed after a quench is not due to barrier crossing but to the emergence of almost flat directions (Castellani & Cavagna, 2005). As explained in (Kurchan & Laloux, 1996), this phenomenon is due to the peculiarity of gradient descent in very high-dimensions; in this case the system is always confined at the border of the basins of attraction, and the Hessian at long times contains a decreasing number of negative eigenvalues, thus leading to an increasingly slow dynamics.

3. Models and Results

We present our core results in two parts: time dependence of the loss function (Sec. 3.1), and identifying different regimes through the two-point correlation function (Sec. 3.2). We start by describing the models used for evaluation:

A - Toy Model: The network contains only 1 hidden layer with $10^4$ hidden nodes. The non-linear function on the hidden layer is ReLU. The output layer is filtered through a sigmoid. The loss function is a mean square error. The total number of weights is around $3 \times 10^8$.

B - Fully Connected: A simple network with three fully connected layers, of sizes 100, 100 and 10, respectively. The non-linear functions are ReLUs, and the loss function is the negative log-likelihood of soft-max outputs. The total number of weights is about $9 \times 10^4$.

C - Small Net: A simple convolutional network with two conv-layers that has 10 and 20 filters in the first and second layer, respectively. It is followed by two fully-connected layers of sizes 100 and 10. The non-linear functions in the hidden layers are ReLUs, and the loss function is the negative log-likelihood of soft-max outputs. The total number of weights is around $6 \times 10^4$.

D - ResNet18: The final model is a ResNet with 18 hidden layers. The total number of weights is around $2 \times 10^7$.

We have chosen networks with various levels of complexity. All networks are initialized in the standard procedures of the PyTorch library (version 0.3.0). The toy model is inspired by the one introduced in (Freeman & Bruna, 2016) which is shown not to have any barriers if the hidden layer is large enough. The training is carried out by SGD that takes a single learning rate that remains unchanged until the end of the computation. The training process runs for a fixed given number of iterations which is deemed to be ‘long enough’ for all practical purposes. For most cases, this means that training kept running long after the perfect accuracy was reached on the training set. All the networks have been trained on multiple datasets: MNIST, CIFAR-10, CIFAR-100, and multiple sets of parameters.

3.1. The Loss Function

We first focus on the time-dependence of the loss function over the training, and we compare it to the one of the energy in glassy systems. For the sake of completeness, we also show the accuracy. We plot the loss values as a function of the logarithm of time, measured in units of iterations so that the unit time step corresponds to a single update of the weights. This choice is different from the wall time or number of epochs which is often used. Although less common in machine learning, the logarithmic scale highlights the slow dynamics and the time dependence. The results obtained for the four networks described above are shown in Figures 2(a), 2(b), 2(c), 2(d). There are several features worth noticing. We can remark three regimes. The first one goes from the beginning of the training up to a time $t_1$, where the loss and accuracy stay roughly constant. At $t = t_1$ the loss starts decreasing roughly linearly in $\log(t)$, and concomitantly the accuracy increases in a similar way. This second regime persists until a time $t_2$, at which the train loss approaches zero. In the final regime beyond $t_2$ the speed of decay sharply decreases. The cross-over times $t_1$ and $t_2$ are indicated in Figures 2(a), 2(b), 2(c), 2(d). In Sec. 3.2 we show that $t_1$ and $t_2$ can also be identified through the evolution of the mean-square displacement.

This behavior is similar to the ones found in disordered systems, see e.g. Figure 1(a). There are however two main differences. First, in several cases the decrease in the second part is actually slower for the DNNs compared to the power-law of the p-spin model. Second, and more importantly, the
Comparing Dynamics: Deep Neural Networks versus Glassy Systems

Figure 2. Train/test loss and accuracy as a function of log(t). The batch size $B$ and learning rate $\alpha$ are specified under each plot. Note that in 2(a) it is more difficult to pin-point the values of $t_1$ and $t_2$ since the crossover is not as sharp as in the other cases.

loss reaches asymptotically (i.e., after $t_2$) its lowest possible value. This is not the case in the $p$-spin model in which instead the energy converges asymptotically to one of the highest and widest minima (Cugliandolo & Kurchan, 1993; Castellani & Cavagna, 2005). Actually, a $p$-spin model with a number of degrees of freedom comparable to the number $M$ of weights that are used in deep learning (in our examples $M = 10^4 - 10^7$) would take an exponentially long time to go beyond the highest and widest minima and reach the bottom of the landscape (Castellani & Cavagna, 2005; Berthier & Biroli, 2011). This is a first indication that the dynamics involved in the training of deep neural networks, although slow, does not correspond to the crossing of large barriers, which would instead lead to much longer time-scales.

In summary, the reason for the slowing down of the dynamics during training is apparently not due to barrier crossing but instead likely related to an increasingly large amount of flat directions that become available to the system during its descent in the loss landscape, as found numerically in (LeCun et al., 1998; Sagun et al., 2017). This is actually similar in the $p$-spin spherical model to the first dynamical regime of aging dynamics that follows a quench. However, in this case the system does not reach the lowest possible values of the loss, as it happens to loss functions during training, but remains trapped in higher and wide local minima.

3.2. Further evidence: Two-time correlation functions

In this section, we focus on the two-time mean-square displacement $\Delta(t_w, t_w + t)$ of the weights and we compare it to the one found for disordered systems after a quench. Its
The results obtained at different \( t_w \) we then normalize the mean-square displacement by \( D(t_w) \), since larger \( D(t_w) \) leads naturally to larger \( \Delta(t_w, t_w + t) \), as illustrated by simple diffusion processes\(^9\).

We present the mean square displacements in Figures 4(a), 4(b), 4(c), 4(d)\(^{10}\). The main result that we find is that for \( t_w < t_2 \) there is a clear \( t_w \) dependence, whereas at larger times the curves for different \( t_w \) collapse together when scaled with \( D(t_w) \). To stress this fact each of the plots has been split in two panels: the upper one shows the curves with \( t_w < t_2 \) and the lower one those with \( t_w > t_2 \)\(^{11}\). The collapse indicates that, except for the change in the strength of the noise \( D \), the dynamics is reaching a stationary regime for \( t_w > t_2 \). In this regime, the loss function is almost equal to zero, thus indicating that the system is diffusing close to the bottom of the landscape\(^{12}\).

Let us now compare this situations with the one of physical systems after a quench, in particular the \( p \)-spin spherical model for \( p = 3 \). In both cases one finds somewhat similar regimes characterized by aging, and corresponding to the descent in the loss (or energy) landscape. The behavior at large times is instead different. In the training dynamics aging is interrupted, meaning that the system becomes stationary except for the change in the noise strength, whereas for the \( p \)-spin model aging persists even when the energy approaches its asymptotic value (on time-scales that do not diverge with the system size). Another difference is the shape of the mean-square displacement curves. During aging, in Figure 1(b), the curves follow a master curve for small \( t \) no matter what is the value of \( t_w \), instead for DNNs no collapse at short-times is present. For \( t_w > t_2 \) the shape of the mean-square displacements does not show any-in

\(^9\)The normalization by \( D(t_w) \) is just an approximate way to take into account the variation of the noise with time; it works well if the variation is not too fast compared to \( t \).

\(^{10}\)For model B and D we averaged over eight and two random initializations, respectively. This is done to iron out the fluctuations of the mean-square displacement. In principle, in order to see the collapse, this procedure should have been carried out for all experiments, but it was not required for models A and C.

\(^{11}\)Except Fig. 4(d), where we could not reach long-enough times, and a hybrid regime is represented.

\(^{12}\)Notes on further experiments: (1) LeNet on CIFAR10 with 77\% test accuracy presents collapse curves at least as good as Figure 4(c), and (2) Deeper ResNet & WideResNet models on both CIFAR10 and CIFAR100 with better accuracies than model D give the correct diffusive slope in their mean square displacement curves but the collapse is not as good as in Figure 4(d). We believe that the key to resolve the collapse in models where number of parameters are much larger goes through a better calculation of the noise coefficient. As a matter of fact, \( D \) changes with time, so rescaling \( \Delta(t_w, t_w + t) \) by \( D(t_w) \) can only work well for small \( t \). This also explains why in Fig. 4 the expected slope \( \Delta / D \sim t \) is only identified for not too large \( t \). We will analyze these issues in detail in an upcoming work.

\[ \Delta(t_w, t_w + t) = \frac{1}{M} \sum_{i=1}^{M} (w_i(t_w) - w_i(t_w + t))^2 \] (3)

where the sum runs over all the weights \( w_i \) of the network, and \( M \) is their total number.

The three regimes of the learning dynamics described in Sec. 3.1 are visible also through the behavior of the mean-square displacement. In Figure 3, for \( t_w < t_1 \), \( \Delta(t_w, t_w + t) \) collapses on a single curve. Once \( t_1 < t_w < t_2 \) the mean-square displacement develops a clear dependence on \( t_w \): the characteristic time increases with \( t_w \), thus showing aging, and when \( t > t_2 - t_w \) it suddenly becomes flat. In the third regime, which corresponds to \( t_w > t_2 \), the characteristic time does not increase any longer with \( t_w \).

To a large extent, the training dynamics at large times can be explained in terms of diffusion in the weight space. A hallmark of a diffusing system is a motion purely driven by the noise \( D \) (Crank, 1979). We estimate the noise in SGD with the variance of the loss function’s gradient\(^8\), which reads (details on the definition of the noise can be found in several resources, see, for example, Li et al. (2015)):

\[ D = \frac{1}{|\text{train set}|} \sum_{s \in \text{train set}} \frac{1}{M} |\nabla \mathcal{L}_s - \nabla \mathcal{L}|^2 \] (4)

where \( \mathcal{L} = \frac{1}{|\text{train set}|} \sum_{s \in \text{train set}} \mathcal{L}_s \) is the empirical average and \( \mathcal{L}_s \) is the loss of the \( s \)-th image in the train set. In a glassy system, the noise is constant through time if the temperature is fixed, whereas during the training \( D \) varies, being a function of the network’s weights. When comparing

![Figure 3. Two-time mean square displacement, \( \Delta(t_w, t_w + t) \), defined in Equation 3, for model C (Small Net). Every curve corresponds to a different waiting time \( t_w \), indicated in the legend.](image)

\[^8\]For reasons of numerical efficiency, for some models \( D \) is calculated on a (sufficiently large) subset of the training set.
Comparing Dynamics: Deep Neural Networks versus Glassy Systems

(a) Toy Model on CIFAR-10, $B = 100$, $\alpha = 0.1$.

(b) Fully Connected on MNIST, $B = 128$, $\alpha = 0.01$.

(c) Small Net on CIFAR-10, $B = 100$, $\alpha = 0.01$.

(d) ResNet-18 on CIFAR-100, $B = 64$, $\alpha = 0.01$.

Figure 4. Mean square displacements rescaled by the noise on the loss’s gradient. Since the behavior of the curves differ in different phases, we show the smaller $t_w < t_2$ on the top set, and the larger $t_w > t_2$ on the lower set. For reference, some $t_w$ appear in both sets. The black segment on the bottom sets represents a slope $\sim t$.

Intermediate plateau\(^{13}\), contrary to what found in Fig. 1(b). The form of $\Delta(t_w, t_w + t)$ is instead the one characteristic of diffusion (the curves $\Delta/D$ would be straight lines in a log-log plot only if $D$ didn’t depend on $t_w$).

Both the aging and the diffusive regimes are present and qualitatively similar in all the analyzed networks. The fact that a slow aging dynamics is also present in model A (Toy Model), that supposedly has no barriers (see Sec. 3), strengthens the conclusion that the dynamics slows down because of the emergence of flat directions that ultimately lead to diffusion at or close to the bottom of the landscape. A deeper analysis of the finer properties of the diffusive regime will be studied in a forthcoming publication.

\(^{13}\)The shape of the mean-square displacements is different for different networks, possibly indicating that the manifolds corresponding to the bottom of the landscape have different geometric characterization.

4. Discussion

In this work we have analyzed the training dynamics of DNNs by methods developed in physics for out-of-equilibrium disordered systems. We have studied the time dependence of the loss value and the mean-square displacements of the weights and compared them to their counterparts in physical systems, in particular the 3-spin spherical spin-glass. The analysis of the time-dependence of the loss function and the mean square displacement indicates that there are at least three time regimes in the training process: one corresponding to an initial exploration of the energy/loss landscape, followed by a decrease of the loss, in which the system displays aging dynamics, and a final regime in which the dynamics appears to be almost stationary and diffusive. Barrier crossing does not seem to play any role. The slowing down can be instead traced back to an increasingly large amount of flat directions that become available to the system.
Comparing Dynamics: Deep Neural Networks versus Glassy Systems

Figure 5. On 5(a) train/test loss and accuracy as a function of log($t$) in a modified version of model A (Toy Model) with only 10 hidden neurons on CIFAR-10. The batch size is $B = 100$, and the learning rate is $\alpha = 0.1$. On 5(b), mean square displacement for the same model.

Figure 5(a). One sees both a collapse at small values of $t_w$, possibly indicating the emergence of an Edwards-Anderson parameter and trapping in bad local minima, and a later $t_w$-dependent time increase just like in regular aging of disordered systems.

On the basis of these results, we conjecture the existence of a phase transition between two regimes: (i) an easy phase corresponding to over-parametrized networks, in which bad local minima do not play any role, dynamics is governed by a massive amount of flat directions, and learning is achieved; (ii) a hard phase corresponding to under-parametrized networks, in which the landscape is rough, dynamics is glassy and the network does not learn well. Whether learning is possible in this case but it would take a huge amount of time to find the good minima is an interesting question.

This scenario has tantalizing similarities with the one found in several combinatorial optimization problems in which easy, hard and impossible algorithmic phases have been found, see e.g. (Monasson et al., 1999; Mézard et al., 2002; Krzakala et al., 2007; Zdeborová & Krzakala, 2016; Achlioptas & Coja-Oghlan, 2008). When degrees of freedom are continuous, the transition between these phases can be associated with the emergence of many flat directions in the energy landscape, a well-known example is the jamming transition of disordered solids (Wyart, 2005; Liu et al., 2010). A detailed investigation of this scenario for DNNs is ongoing and will be presented in a future publication.
Acknowledgements

We thank Valentina Ros for useful conversations. We thank Utku Evci and Uğur Güney for providing the initial version of the code that we used in our numerical simulations. This work was partially supported by the grant from the Simons Foundation (#454935 Giulio Biroli, #454953 Matthieu Wyart, #454951 David Reichman). M.W. thanks the Swiss National Science Foundation for support under Grant No. 200021-165509. M.B.-J. was partially supported through Grant No. FIS2015-65078-C2-1-P, jointly funded by MINECO (Spain) and FEDER (European Union). CC acknowledges support from the Kings Worldwide Partnership Fund.

References


